Высокодоступный MySQL на конвейере

Дмитрий Смаль,
руководитель подразделения
Managed MySQL и SQL Server



Яндекс

MySync. Попробовать прямо сейчас

github.com/yandex/mysync

MySQL в Яндексе

Яндекс Директ, Adfox,
Кинопоиск, Вертикали

Yandex Managed Service
for MySQL®

600 TB
сентябрь 2022

~ 500
кластеров

1000+
хостов

1–15
хостов в кластере

И намного больше кластеров
в Yandex Cloud!

Ожидаемые сценарии

- Учения в дата-центрах
- Плановые работы на гипервизорах
- Изменения и апгрейд кластеров клиентами

И неожиданные проблемки…

- Железо под кластерами ломается
- Сеть также может ломаться
- MySQL полна багов

Чего хотим достичь?

- Доступность 0,9999 в месяц на чтение
- Доступность 0,9995 в месяц на запись
- Спокойный сон дежурного!
- Лёгкая автоматизация



mykote.mirtesen.ru/blog/43046802930

Что нужно автоматизировать?

Обязательно

- Switchover
- Failover
- Переналивка реплики с мастера

Желательно

- Бэкапы с PITR
- Переналивка из бэкапа
- Service Discovery

Какие были решения?

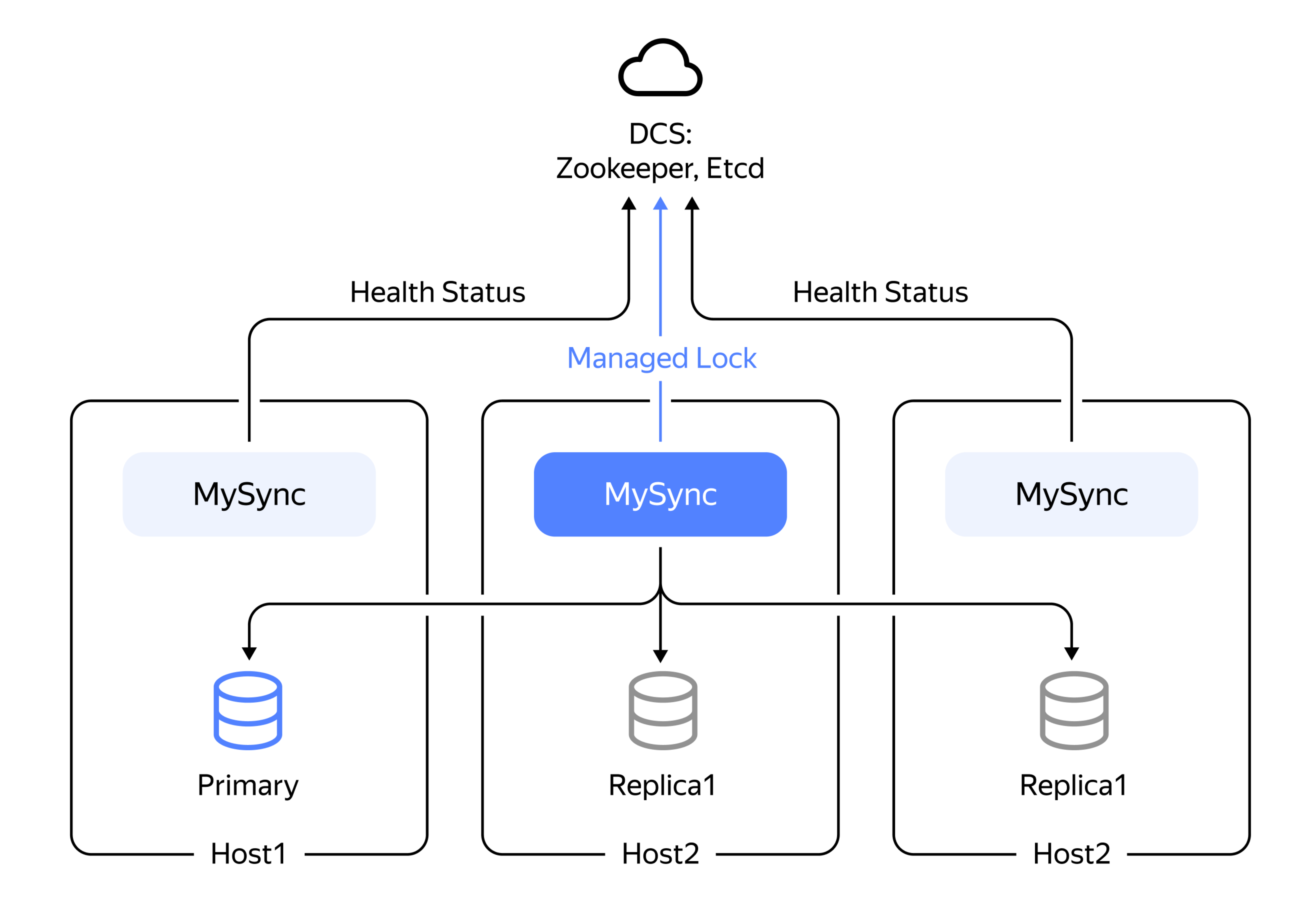
- GitHub Orchestrator
- Severalnines ClusterControl
- MySQL Group Replication
- MHA
- MySync!

MySync.
Архитектура

Основные принципы

- Все хосты эквивалентны
- Идемпотентность операций
- Без потери данных!

Архитектура MySync



Процесс переключения мастера

Что должна сделать HA-утилита?

- Перевести кластер в RO
- Остановить репликацию
- Выбрать лучшую реплику (*)
- Повернуть все реплики на неё
- Открыть новый мастер
- Обновить записи в DCS/SD

Проверка состояния мастера

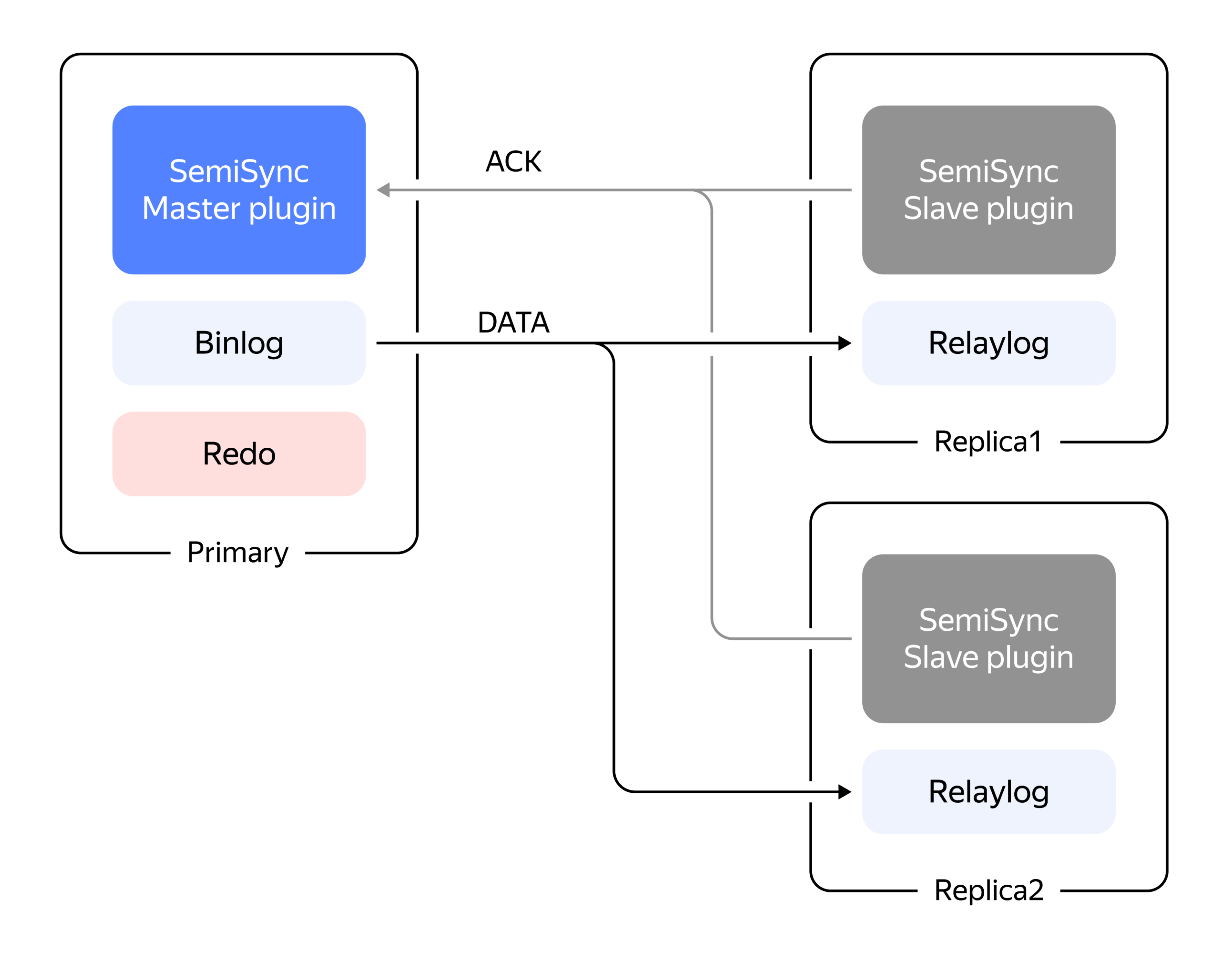
Подход в MySync

- Наличие соединения с DCS
- Сервер отвечает на SELECT 1

Но есть 99 способов ошибиться

- Файловая система в RO
- Исчерпание ресурсов
- Проблемы с DCS

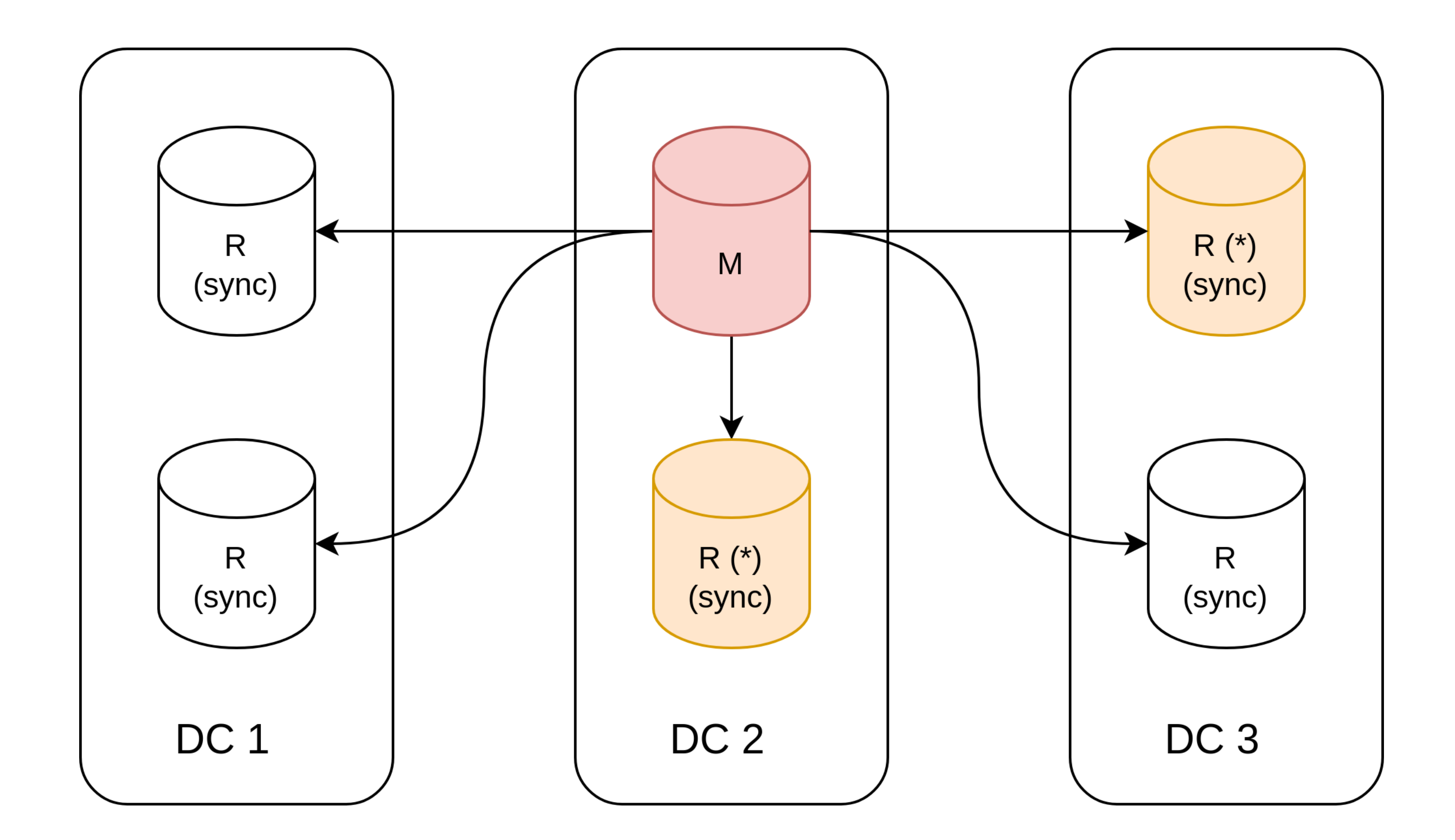
Синхронная репликация в MySQL



Кворум

W + R = N + 1

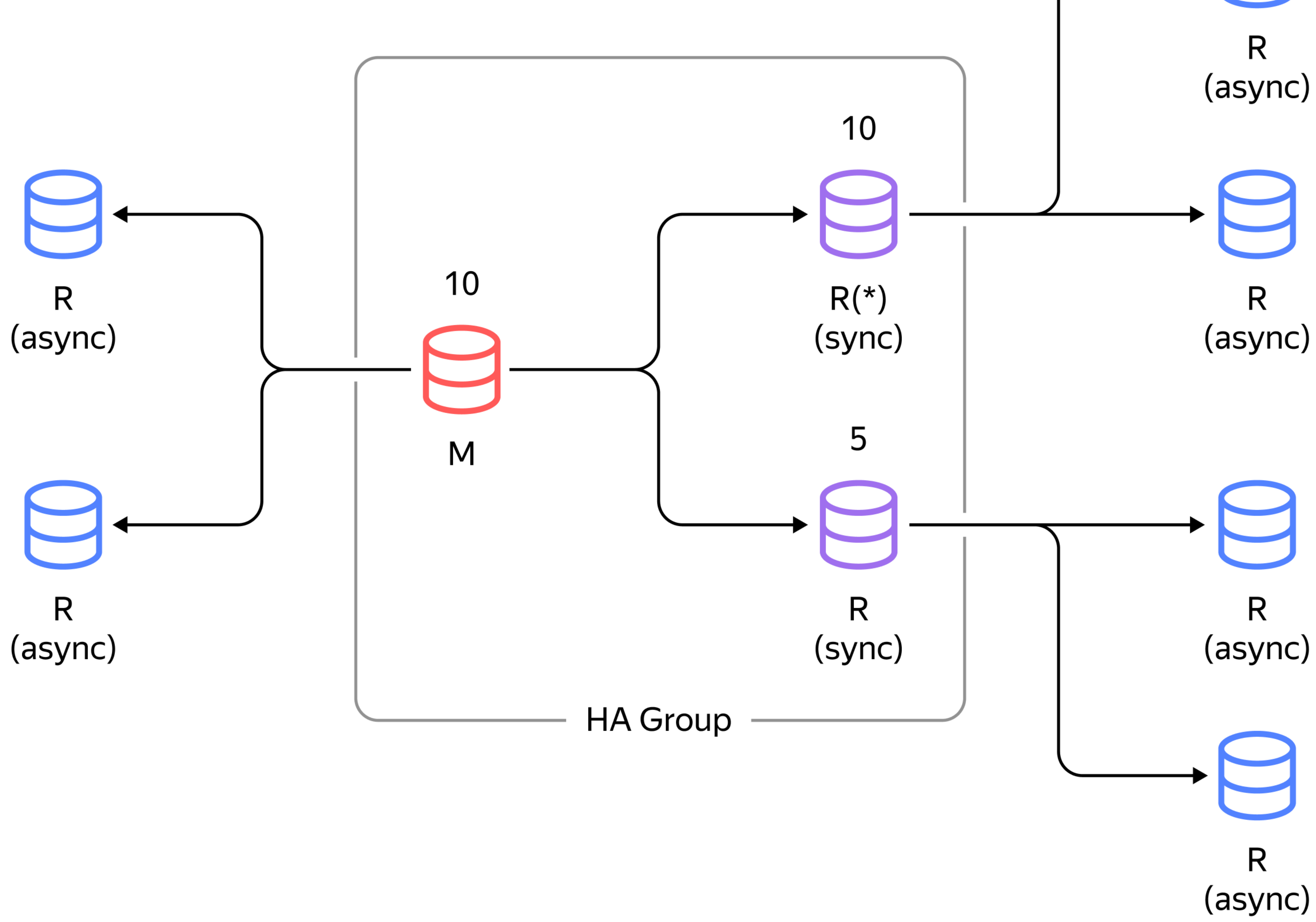
- N — число хостов в кластере
- W — число хостов, на которых была записана транзакция
- R — сколько хостов должно выжить



В MySQL

- W = 1 + rpl_semi_sync_master_wait_for_slave_count

Каскадная репликация и приоритеты



Режим maintenance

В режиме maintenance

- отключает синхронную репликацию
- не делает никаких изменений

Нужно

- для ручных работ на кластере
- в процессе изменения кластера



Переналивка реплик

MySync

- Проверяет необходимость touch/var/run/mysync/mysync.resetup

Внешняя автоматика

- Проверяет наличие файла
- Переналивает инстанс mysql
- удаляет mysync.resetup



npfgeoprom.ru/bezopasnost/ne-vlezaj-ubet-znak.html

Что имеем в результате

Если всё идет хорошо

~10–20 с

Switchover

~1–2 мин

Failover

- Без потери данных
- Последовательная деградация N → 1 хост

Как всё испортить?

Что плохого может сделать пользователь СУБД ?

- Большие транзакции
- Отстающие реплики
- Ненадёжная конфигурация MySQL

Но мы будем разумными

Best practices

- Правильные индексы в таблицах
- Небольшие транзакции
- Pt-online-schema-change

Настройки MySQL

- slave_rows_search_algorithms = INDEX_SCAN,HASH_SCAN
- slave_parallel_type = LOGICAL_CLOCK
- slave_parallel_workers = 8

MySync

Как попробовать?

Скачать

```
git clone git@github.com:yandex/mysync.git
```

Собрать

```
go build -o mysync ./cmd/mysync/...
```

Настроить MySync

```
failover: true
failover_cooldown: 3600s
failover_delay: 60s
zookeeper:
  namespace: /mysql/mycluster_1
  hosts:
    - zk-dbaas02f.db.yandex.net

    - zk-dbaas02h.db.yandex.net

    - zk-dbaas02k.db.yandex.net
mysql:
  user: admin
  password: ******
  port: 3306
  replication_port: 3306
  replication_user: repl
  replication_password: ******
```

Настроить MySQL

репликация

```
binlog_format = ROW
gtid_mode = ON
enforce_gtid_consistency = ON
log_slave_updates = ON
```

сохранность данных

```
innodb_flush_log_at_trx_commit = 1
sync_binlog = 1
```

избегаем split-brain

```
read_only = ON
super_read_only = ON
offline_mode = ON
```

Использовать

```
# собрать первоначальную конфигурацию руками
# сообщить MySync о хостах
$ mysync host add  node1.db.your.project.com
$ mysync host add  node2.db.your.project.com
$ mysync host add  node3.db.your.project.com

# посмотреть информацию
$ mysync info -s

# переключить мастер
$ mysync switch --to node2
$ mysync switch --from node2

# maintenance
$ mysync maint on|off
```

Что ещё умеет MySync?

- «Починка» кластера
- Управление свободным местом
- Закрытие отстающих реплик



victorinox.ru/

Какие планы развития?



victorinox.ru/

- Поддержка каналов репликации
- Поддержка MariaDB
- Поддержка Etcd
- LibMySync, API?



Дмитрий Смаль

Руководитель подразделения Managed MySQL и SQL Server

mialinx@yandex-team.ru



Обратная связь
и комментарии
к докладу по ссылке





HighLoad++ 2022

Яндекс